From Physics to Algorithms: New Approaches to Information Science

1. PI: Michael Chertkov (T-13) #149600

2. Additional Investigators:

- Frank Alexander (CCS-3) #110216
- Eli Ben-Naim (T-13) #119658
- Brent Daniel (D-3) #180380
- Anders Hansson (CCS-5) #191558
- Matthew Hastings (T-13) #175388
- Gabriel Istrate (CCS-5) #169454
- Allon Percus (CCS-3) #146353
- Charles Reichhardt (T-13) #170260
- Mikhail Stepanov (T-13) #188972

3. Primary DOE Mission addressed

- **Reducing the Treats From Weapons of Mass Destruction and Advancing National Defense Capabilities**

4. Additional Missions Supported by Work

- **Enhancing the Nation's Fundamental Scientific Capabilities**
- **Improving the Nation's Energy Security**

5. Support for either or both of the Capability Thrusts

- **Predictive Science**
- **Materials Science**

6. Budget request for FY07, FY08, and FY09, in $K

- FY07: $1200K
- FY08: $1250K
- FY09: $1300K

Science and Technology Objectives

The information revolution has dramatically raised the standards for reliable transmission, storage, and processing of information. This requires new, efficient algorithms for computationally hard problems. Many practical methods have been developed on a heuristic level, without a theoretical basis, which severely limits their scope of applicability. To bridge the gap between engineering needs and the traditional computer science rigorous approaches, we propose an innovative physics-inspired program for the *analysis and development of algorithms*. The goal is to reduce the computing time and memory space required for complex operations in information processing.

There are two crucial metrics for algorithmic performance: (1) the quality of the solution found and (2) the resources (computing time and memory space) required to find the solution. Algorithms that guarantee the best solution to computationally hard problems require exponentially large time, while algorithms that allow near-optimal solutions can run much faster. We will consider the typical performance of an algorithm, as well as the worst-case performance for extreme scenarios where high reliability is required.

We will develop methods that work well in practice. Data sets on personal computers can easily involve 10^{12} bits, while supercomputing and databank applications involve much larger data sets. The techniques of statistical physics, developed to study macroscopic systems of 10^{23} particles, offer a natural approach to dealing with such sizes. Moreover, statistical physics has a well-developed methodology for describing problems of large but finite size. These issues are becoming especially important in modern physics applications at the nanoscale and are now being imported into information and computer sciences, where finite-size performance is obviously crucial.

Algorithm analysis. We will exploit powerful recent developments at the intersections of theoretical physics, information theory, and computer science to analyze the performance of existing algorithms, such as the *belief propagation* algorithm that is broadly used in the three fields. We will study algorithms for data transmission, storage and retrieval problems in information science. In addition, we will analyze fundamental combinatorial issues of computer science that arise in verification and validation and contingency planning problems. The best approximation algorithms find solutions that are close to optimal in all but a few cases. For example, the belief propagation algorithm fails when loops in the underlying graphical structures become important. We will analyze these cases using powerful methods of theoretical and mathematical physics. Additionally, we will use phase transition methods such as scaling, renormalization, and universality to characterize algorithms that work well in certain regimes but perform poorly in others.

Algorithm development. Many information science problems can be mapped onto physical systems of interacting particles or spins. Our approach is to find such mappings and then utilize them for the design of new algorithms. We will also develop a novel rational approach which is built on recasting optimal but computationally intractable algorithm in a series of systematically improvable approximations. The higher the order of an approximate algorithm in the sequence the better its quality is expected to be while the respective computational complexity will still be kept polynomial. Theoretical analysis and large scale computation will be used to quantify this improvement.

Tasks and Probable accomplishments

The proposed research detailed below is organized into two complementary categories:

Information Science. Data storage and retrieval is a classic area of information science. Handling massive amounts of data requires compression, decompression, storage and then retrieval of data.

We will develop methods for first characterizing and then qualitatively improving the performance of data storage and retrieval algorithms. Many data storage and sorting algorithms currently utilize tree architectures. We propose the novel approach of mapping trees to collision processes in a gas, using nonlinear dynamics and stochastic processes techniques. Our goal is to address how data retrieval protocols scales with the size of the data set. In addition, we will analyze modern data retrieval processes that utilize natural architectures such as a *two-dimensional lattice*, instead of the traditional co-centric architecture, or the aforementioned tree architecture. This emerging technology, relevant to high density recording, presents a challenge: overcoming data corruption that may occur in the reading process. Luck of efficient algorithms is the current bottleneck in the high-density data restoration and related areas of two-dimensional and three-dimensional (holographic) image processing. Mapping the multi-dimensional inference problem to a spin system will allow us to develop efficient algorithmic solutions for the structured data restoration and image processing.

Another major theme is coding. In a general coding scheme the transmitter wishes to send data over a noisy channel and so encodes the data in a redundant form, sending a longer message over the channel. The receiver gets a corrupted message, and then tries to reconstruct the transmitter's original message. The practical decoding challenge is to reconstruct the best possible approximation to original data with minimal computational complexity. We have recently developed a comprehensive theoretical and computational framework for describing the performance of error correcting codes and we will use this framework to design improved decoding algorithms. Next, we will analyze and develop algorithmic ideas in the broader context of distributed systems. A new development is that of *network coding* – an emerging field taking a unified approach to previously segmented problems, like coding and routing, related to information processing between multiple peers. This research will target applications in detection of biological agents and toxic chemicals, environmental measurement of radiation, and real-time video surveillance.

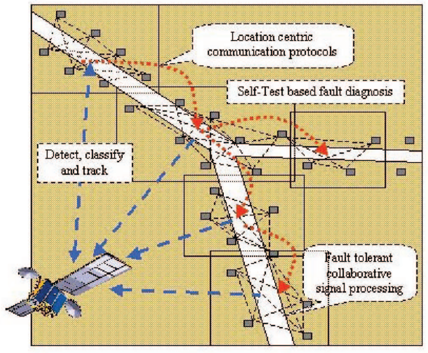
Location centric communication protocols
Self-Test based fault diagnosis
Detect, classify and track
Fault tolerant collaborative signal processing

Figure 1: Wireless Sensor Network – complex real-world problem with algorithmic challenges in coding, secure routing and energy efficient clustering.

<u>*Computer Science.*</u> We will study two important challenges in computer science: satisfiability and community detection. *Satisfiability* is a canonical computer science problem that promises significant contributions to the Laboratory's verification and validation efforts. In satisfiability, given a set of logical clauses, values must be assigned to Boolean variables to satisfy all (or a maximal number of) clauses. While this problem is in general intractable, certain heuristic algorithms can work very well on typical instances of it. Recently, an understanding of this situation has come about through statistical physics models that associate the "intractable" cases with phase transitions. This understanding is leading to the *design* of effective algorithms, both at LANL and elsewhere, for solving some of these hardest instances. Doing so is of very practical concern. One of the hottest applications of satisfiability is in *formal verification*: rigorously proving the correctness of computer programs. (Microsoft has an outstanding research group applying statistical physics and computer science to formal verification of device drivers.) We plan to apply formal verification techniques to complex scientific codes such as ocean models and multiphysics problems, where algorithms such

as adaptive mesh refinement pose significant challenges to more conventional verification methods. We believe that modern satisfiability solvers will improve substantially upon the state-of-the-art in verification of advanced scientific codes, and the needs of these codes will simultaneously drive further theoretical development of satisfiability algorithms.

The other fundamental problem we intend to address, *community detection*, involves decomposing a graph or relational database into communities such that nodes within a community are highly connected to each other but not to other communities. The problem is closely related to graph partitioning, and is of strong interest for graphs found in social and biological systems. Existing algorithms, while in many cases successful at revealing hidden structures in networks, use largely ad hoc definitions of community. We plan to approach the challenge of community detection in a different way, formulating it as an inference problem and picking the community assignment with the maximal likelihood of leading to the final graph. In cases of practical interest, the number of communities is small compared to the number of nodes: this scenario is ideally suited to the framework of *belief propagation*. We expect that the systematic basis of belief propagation methods will lead to significantly improved performance compared to current methods. Furthermore, belief propagation is very rapid, and offers the prospect of solving the community detection problem on massive data sets encountered in biological and social systems. These are essentially beyond the reach of present algorithms.

Institutional Goals and Objectives

There is a need for the United States to maintain its leadership in the field of information technology. It is scientifically unhealthy and strategically unwise to rely on foreign countries to provide the strategic technologies and knowledge base critically needed in defense and national security computing applications. Our proposal is one step in the development of a long-term capability at Los Alamos in the increasingly pervasive area of information science. Specifically, the development of efficient algorithmic solutions for computationally hard tasks is fundamental to the Laboratory's mission. It impacts programs in modeling and simulation, verification and validation, nonproliferation, homeland and infrastructure security.

Our proposed research fits into a long-standing tradition at the Laboratory. From MANIAC to the introduction of the Monte Carlo method, to the creation of the scholarly information archive (xxx.lanl.gov), there has been a lively institutional dynamic that involves leveraging our strengths in the physical sciences to secure leadership in information and computer science. Los Alamos has made a strategic decision to invest in theoretical statistical physics for national security applications, and this investment is paying off with visionary programmatic work, high-impact basic research, and recruitment of outstanding young talent.

Our team includes theoretical physicists and information and computer scientists, who are already working in close collaboration. This interdisciplinary group is highly visible in the international research community. Our team has a strong track record in analyzing and developing algorithms using physical principles. Key accomplishments include transmission fidelity analysis for optics communications and error analysis in coding algorithms using field theory, traveling wave techniques for analysis of data storage, compression and zipping algorithms, and development of the extremal optimization algorithm used successfully on notoriously intractable problems. (For references and additional information, see http://cnls.lanl.gov/˜chertkov/alg.htm.)